Abstract

We consider supervised learning problems, where one is given objects with labels, and the goal is to learn a model that can make accurate predictions on new objects. These problems abound in applications, ranging from medical diagnosis to information retrieval to computer vision. Examples include binary or multiclass classification, where the goal is to learn a model that can classify objects into two or more categories (e.g. categorizing emails into spam or non-spam); bipartite ranking, where the goal is to learn a model that can rank relevant objects above the irrelevant ones (e.g. ranking documents by relevance to a query); class probability estimation (CPE), where the goal is to predict the probability of an object belonging to different categories (e.g. probability of an internet ad being clicked by a user). In each case, the accuracy of a model is evaluated in terms of a specified 'performance measure'.

While there has been much work on designing and analysing algorithms for different supervised learning tasks, we have complete understanding only for settings where the performance measure of interest is the standard 0-1 or a loss-based classification measure. These performance measures have a simple additive structure, and can be expressed as an expectation of errors on individual examples. However, in many real-world applications, the performance measure used to evaluate a model is often more complex, and does not decompose into a sum or expectation of point-wise errors. These include the binary or multiclass G-mean used in class-imbalanced classification problems; the F_1-measure and its multiclass variants popular in text retrieval; and the (partial) area under the ROC curve (AUC) and precision@κ employed in ranking applications. How does one design efficient learning algorithms for such complex performance measures, and can these algorithms be shown to be statistically consistent, i.e. shown to converge in the limit of infinite data to the optimal model for the given measure? How does one develop efficient learning algorithms for complex measures in online/streaming settings where the training examples need to be processed one at a time? These are questions that we seek to address in this thesis.

Firstly, we consider the bipartite ranking problem with the AUC and partial AUC performance measures. We start by understanding how bipartite ranking with AUC is related to the

standard 0-1 binary classification and CPE tasks. It is known that a good binary CPE model can be used to obtain both a good binary classification model and a good bipartite ranking model (formally, in terms of regret transfer bounds), and that a binary classification model does not necessarily yield a CPE model. However, not much is known about other directions. We show that in a weaker sense (where the mapping needed to transform a model from one problem to another depends on the underlying probability distribution), a good bipartite ranking model for AUC can indeed be used to construct a good binary classification model, and also a good binary CPE model. Next, motivated by the increasing number of applications (e.g. biometrics, medical diagnosis, etc.), where performance is measured, not in terms of the full AUC, but in terms of the partial AUC between two false positive rates (FPRs), we design batch algorithms for optimizing partial AUC in any given FPR range. Our algorithms optimize structural support vector machine based surrogates, which unlike for the full AUC, do not admit a straightforward decomposition into simpler terms. We develop polynomial time cutting plane solvers for solving the optimization, and provide experiments to demonstrate the efficacy of our methods. We also present an application of our approach to predicting chemotherapy outcomes for cancer patients, with the aim of improving treatment decisions.

Secondly, we develop algorithms for optimizing (surrogates for) complex performance measures in the presence of streaming data. A well-known method for solving this problem for standard point-wise surrogates such as the hinge surrogate, is the stochastic gradient descent (SGD) method, which performs point-wise updates using unbiased gradient estimates. However, this method cannot be applied to complex objectives, as here one can no longer obtain unbiased gradient estimates from a single point. We develop a general stochastic method for optimizing complex measures that avoids point-wise updates, and instead performs gradient updates on mini-batches of incoming points. The method is shown to provably converge for any performance measure that satisfies a uniform convergence requirement, such as the partial AUC, precision@κ and F_1-measure, and in experiments, is often several orders of magnitude faster than the state-of-the-art batch methods, while achieving similar or better accuracies. Moreover, for specific complex binary classification measures, which are concave functions of the true positive rate (TPR) and true negative rate (TNR), we are able to develop stochastic (primal-dual) methods that can indeed be implemented with point-wise updates, by using an adaptive linearisation scheme. These methods admit convergence rates that match the rate of the SGD method, and are again several times faster than the state-of-the-art methods.

Finally, we look at the design of consistent algorithms for complex binary and multiclass measures. For binary measures, we consider the practically popular plug-in algorithm that constructs a classifier by applying an empirical threshold to a suitable class probability estimate,

and provide a general methodology for proving consistency of these methods. We apply this technique to show consistency for the F_1-measure, and under a continuity assumption on the distribution, for any performance measure that is monotonic in the TPR and TNR. For the case of multiclass measures, a simple plug-in method is no longer tractable, as in the place of a single threshold parameter, one needs to tune at least as many parameters as the number of classes. Using an optimization viewpoint, we provide a framework for designing learning algorithms for multiclass measures that are general functions of the confusion matrix, and as an instantiation, provide an efficient and provably consistent algorithm based on the bisection method for multiclass measures that are ratio-of-linear functions of the confusion matrix (e.g. micro F_1). The algorithm outperforms the state-of-the-art SVMPerf method in terms of both accuracy and running time.

Overall, in this thesis, we have looked at various aspects of complex performance measures used in supervised learning problems, leading to several new algorithms that are often significantly better than the state-of-the-art, to improved theoretical understanding of the performance measures studied, and to novel real-life applications of the algorithms developed.